# Hypergeometric Distribution Based Semantic Searching Technique

N Thakur
Asst. Prof, Dept. of Computer Science & Engg
Bharati Vidyapeeth's College of Engineering
New Delhi, India
narinat@gmail.com

S Gupta
Dept. of Computer Science & Engg.
Bharati Vidyapeeth's College of Engineering
New Delhi, India
siddharth_bvcoe@hotmail.com

## ABSTRACT

Semantic Web is, without a doubt, gaining momentum in both industry and academia. The word "Semantic" refers to "meaning" – a semantic web is a web of meaning. A web that knows what the entities on the web mean can make use of that knowledge. Why do we think that the web would be improved, if it understood the meaning of its contents? Doesn't it understand it now? Google is very good at correcting typing mistakes, figuring out what I "meant" when I miss-typed a query or suggesting keyword to expand our search query. In this paper we redefine the idea of searching related text information on web. The syntactical characters of related keywords and texts are described in detail, but it doesn't involve semantics of the keywords. Hence computers are able to determine the related keywords and texts without actually understanding the meanings or relevance of the keywords. Here Hypergeometric distribution model is used which is a discrete probability distribution that describes the number of successes in a sequence of n keywords to be searched from a finite population without replacement.

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**] : Information Filtering, Query Formulation, Retrieval Models, Search Process

## General Terms

Algorithms, Performance, Experimentation, Languages, Theory

## KEYWORDS

Hypergeometric Distribution, Searching, Optimisation, Capture-Recapture Method

## 1. INTRODUCTION

The Semantic Web is an extension of the World Wide Web with new technologies and standards that enable interpretation and processing of data and useful information for extraction by a computer. The Web contains a huge amount of data but computers alone cannot understand or make any decisions with this data [5]. The solution is the Semantic Web. The Semantic Web is not a separate web but an extension of the World Wide Web, in which information is given well defined meaning, better enabling computers and people to work in cooperation. [4] Sometimes it is said that the Semantic Web will make data become "smart". What would it mean, for data to be smart? Smart data means that the web of information becomes so richly interconnected that it can become much smarter than humans [10]. With the emergence of Semantic Web framework the naïve approach of searching information on the syntactic web is cliché. The World Wide Web is a congregation of billion web pages which are adhered to each other through hyperlinks. Many internet users daily activity is web search only and these users end up in an endless quench to retrieve relevant information pertaining to the user in shortest possible time. A Web 2.0 based search engine is having major drawback as it lacks interpretability between machines, metadata and knowledge management crisis [11]. Powerful and complex algorithms are required by the search engines in order to parse the keywords requested by the user. The future web, Semantic Web is based on the principal of interoperability between machines and giving them power to think [10][11] which aims at attaching metadata, specifying relations between web resources and knowledge management, in order to process and integrate data by the users.

In our present syntactic web, the searching is keyword based where when we attempt to find information for a search query, all the keyword sets are obtained related to that web page which tells us about the related information but if the typed words are not in the keyword sets, computer will not give it to us [13][14].

Our proposal in this paper redefines the relevance of keywords by using the syntactical characters of the keywords and applying the Hypergeometric distribution function which enables us to determine the relevance of the keywords. For instance, let a user types a keyword set X. Then by looking at the texts found by means of X, some other words related to X can be determined like antonyms and synonyms. Obviously, these texts may not include any query word and they include merely one related word or more [9][12]. The Hypergeometric distribution aims at finding the probabilities of the keywords with the maximum likelihood occurrence if keywords in a set. We have also applied Capture-Recapture method which will help us in assigning priority to the web pages based on the high probability.


*ICWET'11*, February 25–26, 2011, Mumbai, Maharashtra, India.

We thank our family members and colleagues for supporting us throughout in our research.

## 2. BACKGROUND STUDY

The Semantic Web which is in the early stages of development has not reached the stage of pervasive web of distributed knowledge and searching based on intelligence. Success will be attained when a dynamic synergism can be created between people and a sufficient number of infrastructure systems and tools for the Semantic Web in analogy with those for the original web.

Numerous parallel research works has been done in order to efficiently retrieve data from the web. Samantha K. Rajapaksha in [11] talks about semantic query model in which he takes list of plain keywords as inputs and equivalent semantic queries are expressed in knowledge representation language. Semantic Web is a web of databases and not of documents, queried by SPARQL [6]. RDF attaches metadata specify relations between the resources based on XML. Ontologies are another major Semantic Web technology built above RDF aims at providing strong semantics and vocabulary [7] [8]. They link the data on the basis of logical reasoning, common vocabulary and analytical thinking.

Another interesting approach has been devised by Qing Zhou and ZeQI Zheng in [9], where intelligent expansion of query of searching related text information is proposed. His idea of proposal was: Let user types a search query K. Then by looking at the texts found by means of K, other words related to K can be determined. Example if we type "database" as our search query then the computer will give us all the texts on the web containing the word "database" in their keysets. Now these keysets will also have other words related to "database". As the keywords of the same set are interrelated to each other, he proposed an algorithm to search related text information by searching those which have keywords in the keyword sets of the texts we already have [9][12].

The closest keyword here is "Conference" which is occurring most number of times in the keysets. The above algorithm illustrates the semantic searching of words on the web. As we know present scenario of searching depends on heavy parsing algorithms in order to yield results as computer cannot understand the meaning of the documents. Here we have retrieved all the keysets related to the search query in the database and the union of all the keysets are taken in another set. Our method, exemplifies the retrieval of keywords by finding out the Hypergeometric distribution function of the keywords obtained in the union keyword sets, by which we can assign priority to the pages which needs to be addressed first to the user containing higher probability keywords in their keyword sets.

## 3. HYPERGEOMETRIC DISTRIBUTION

A Hypergeometric random variable is the number of successes that result from a Hypergeometric experiment. The probability distribution of a Hypergeometric random variable is called a Hypergeometric distribution. Given *x*, *N*, *n*, and *k*, we can compute the Hypergeometric probability based on the following formula:

According to the Hypergeometric Formula [14], let us suppose a population consists of *N* items, *k* of which is successes and a random sample drawn from that population consists of *n* items, *x* of which are successes. Then the Hypergeometric probability is:

$$h(x,\ N, n, k) = [\ {}_{k}C_{x}\ ]\ [\ {}_{N-k}C_{n-x}\ ]\ /\ [\ {}_{N}Cn\ ]$$

The Hypergeometric distribution has the following properties:

- The mean of the distribution is equal to $n * k / N$.
- The variance is $n * k * (N - k) * (N - n) / [N2 * (N- 1)]$.

Example1

Suppose we randomly select 5 cards without replacement from an ordinary deck of playing cards. What is the probability of getting exactly 2 red cards (i.e., hearts or diamonds)?

*Solution:* This is a Hypergeometric experiment in which we know the following:

- N = 52; since there are 52 cards in a deck.
- k = 26; since there are 26 red cards in a deck.
- n = 5; since we randomly select 5 cards from the deck.
- x = 2; since 2 of the cards we select are red.

We plug these values into the Hypergeometric formula as follows:

$$h(x; N, n, k) = [\ {}_{k}C_{x}\ ]\ [\ {}_{N-k}C_{n-x}\ ]\ /\ [\ {}_{N}C_{n}\ ]$$
$$h(2; 52, 5, 26) = [\ {}_{26}C_{2}\ ]\ [\ {}_{26}C_{3}\ ]\ /\ [\ {}_{52}C_{5}\ ]$$
$$h(2; 52, 5, 26) = [\ 325\ ]\ [\ 2600\ ]\ /\ [\ 2{,}598{,}960\ ] = 0.32513$$

Thus, the probability of randomly selecting 2 red cards is 0.32513.

In this paper 'w balls' are drawn randomly from the bag is related to the randomly taken related/unrelated information/pages/links from the web and among these links the relevant links/web pages are passed on and all these related/unrelated links/pages(return all the w balls to the bag) are returned to web. The related links will be passed on as a result on the priority basis by using the maximum likelihood method as discussed below. Here we establish a way to estimate numerical values of parameters m and w from the observed x (k)'s. The method which we propose for searching of related text information is maximum likelihood

The likelihood function l (m, w) is evaluated based on the Hypergeometric distribution as:

$$l(m,w) = \prod_{i=1}^{n} Prob(x(i)|m,w,C(i-1))$$

where n is the total number of x (i)'s.

Thus the condition of the maximum likelihood can be formulated as follows, ignoring for a while that the combinatorial in (1) are discrete functions and therefore their differentials are not defined in the strict sense [3].

$$l(m,w) = \prod_{i=1}^{n} Prob(x(i)|m,w,C(i-1))$$

Here Capture-Recapture method is used which is one of the methods for estimating the size of wildlife populations and is based on the Hypergeometric distribution. A Hypergeometric distribution is a three-parameter family of discrete distributions and one of the parameters, denoted by ‘N’ in this post, is the size of the population. We show that the estimate for the parameter ‘N’ that is obtained from the Capture-Recapture method is the value of the parameter ‘N’ that makes the observed data “more likely” than any other possible values of ‘N’. Thus, the Capture-Recapture method produces the maximum likelihood estimate of the population size parameter ‘N’ of the Hypergeometric distribution.

Let’s start with an example. In order to estimate the number of the web pages/ links which contain the keywords (as in search query “IEEE”) in the web, a total of w= 250 keyset are found/captured and tagged and then released.

After allowing sufficient time for the tagged keywords, a sample of n =150 web pages/ links were found/caught which contain the query keyword. It was found that y= 16 web pages/ links were tagged. Estimate the size/ number of the web pages/ links in the web.

Let ‘N’ be the size of the web pages/ links which contain the keywords in the web. The population proportion of the tagged web pages/ links which contain the keywords is 16. The sample proportion of the tagged web pages/ links which contain the keywords is (y/n). In the capture-recapture method, the population proportion and the sample proportion are set equalled. Then we solve for ‘N’.

$$\frac{w}{N} = \frac{y}{n} \Rightarrow N = \frac{wn}{y} = \frac{250(150)}{16} = 2343.75 = 2343$$

Now, let us related this to the Hypergeometric distribution. After w=250 web pages/ links which contain the keywords were captured, tagged and released, the population is separated into two distinct classes, tagged and non-tagged. When a sample of n = 150 web pages/ links which were selected without replacement, we let ‘Y’ be the number of web pages/ links which contain the keywords in the sample that were tagged. The distribution of ‘Y’ is the Hypergeometric distribution. The following is the probability function of ‘Y’.

$$P\,[Y = y] = \frac{\binom{w}{y}\binom{N-w}{n-y}}{\binom{N}{n}}$$

In the Hypergeometric distribution described here, the parameters w and n are known (w=250 and n= 150). We now show that the estimate of N= 2343 is the estimate that makes the observed value of y=16 “most likely” (i.e. the estimate of N= 2343 is a maximum likelihood estimate of N). To show this, we consider the ratio of the Hypergeometric probabilities for two successive values of N.

$$\frac{P(N)}{P(N-1)} = \frac{(N-w)(N-n)}{N(N-w-n+y)}$$

where $P(N) = \frac{\binom{w}{y}\binom{N-w}{n-y}}{\binom{N}{n}}$ and $P(N-1) = \frac{\binom{w}{y}\binom{N-1-w}{n-y}}{\binom{N-1}{n}}$

Note that $1 < \frac{P(N)}{P(N-1)}$ or $P(N-1) < P(N)$ if and only

if the following holds:

$N\,(N - w - n + y) < (N - w)\,(N - n)$

$N < \frac{wn}{y}$

Note that (wn / y) is the estimate from the capture-recapture method. It is also an upper bound for the population size N such that the probability P (N) is greater than P (N-1). This implies that the maximum likelihood estimate of N is achieved when the estimate is equal to $\frac{wn}{y}$ . As an illustration, we compute the probabilities

$$P(N) = \frac{\binom{250}{16}\binom{N-250}{150-16}}{\binom{N}{150}}$$

for several values of N above and below N=2343. The following matrix illustrates that the maximum likelihood is achieved at N=2343.

$$\begin{pmatrix} N & P(N) \\ 2340 & 0.1084918 \\ 2341 & 0.1084929 \\ 2342 & 0.1084935 \\ 2343 & 0.1084938 \\ 2344 & 0.1084937 \\ 2345 & 0.1084933 \\ 2346 & 0.1084924 \end{pmatrix}$$

Considering the sampled data and using Hypergeometric formulae the various Hypergeometric distributions (chart) are drawn as: Let the total number of keyword i.e. Lot size N = 2343, Keywords which are matching the query i.e. Successes of lot M = 250, we found a sample without replacement of n keywords i.e. the Sample size n= 150 and let x (Successes of sample x) equal the number of keywords in our sample of size be 16. We draw a Hypergeometric Distribution graph indicating the distribution function and probability mass function of the number of successes.

## 4. RESULT AND ANALYSIS

Probability Function

$$f(x, n, M, N) = \frac{{}_{M}C_{x} \quad {}_{N-M}C_{n-x}}{{}_{N}C_{n}}$$

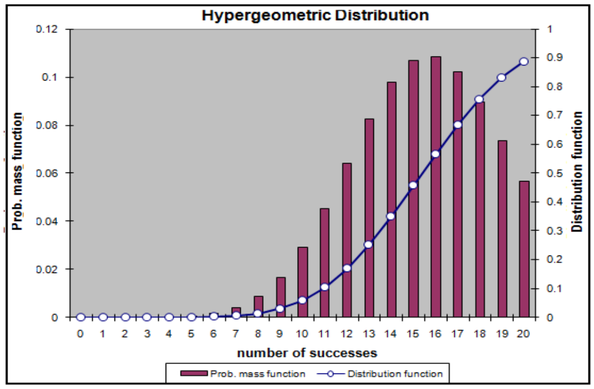


Fig. 2 Hypergeometric Probability graph

Let's start with the above example considering:

**Search Query= {"IEEE", "India", "Technology"}**

Now, union of all the key sets are taken, say
P = {"IEEE", "Conference", "Indian", "Universities", "Technology", "Society"}

6 keysets —3 main keywords and 3 are relative keywords . There is a total of 6 keywords. Combination of 6 taken 2 at a time is $^{3\times 2}C_2 = {}^{6}C_2 = 15$. The following two restrictions apply:

1) No 2 relative keywords at a time is a valid pairing;
2) No 2 main keywords at a time is a valid pairing.

Combinations of 3 main keywords taken 2 a time is $^{3}C_2 = 3$. Combinations of 3 relative keywords taken 2 a time is $^{3}C_2 = 3$. The total number of exclusions is 2 x 3 = 6. If we deduct the exclusions from total number of combinations, the result is: 15 – 6 = 9. The result represents the square of the number of key sets: $3^2$.

The 3 key sets consist of {Main1+relative1}, {Main 2+ relative 2}, {Main + relative 2}. These are the 9 pairings:
M1+R1, M1+R2, M1+R3
M2+R1, M2+R2, M2+R3
M3+R1, M3+R2, M3+R3

There are 3 cases of equal index: M1+R1, M2+R2, M3+R3. They represent the unfavourable cases. The remaining 9 – 3 = 6 key sets represent the cases of interest:
We can generalize for a number of N key sets.
$^{2N}C_2 = \{2xN \times (2N-1)\} / \{1 \times 2\} = \{N \times (2xN - 1)\} = 2N^2 - N$.

*Total possible cases =* $N^2$.

## 5. CONCLUSION

We have addressed a Hypergeometric distribution and Capture-Recapture method for estimating the Key sets probabilities in Symantec Searching Technique. Here we have proposed an algorithm for Web Searching technique by assigning priority to the web pages based on the high probabilities of the Search Keysets. The advantage of using this technique is that one can prioritise the occurrence of web pages and semantically search the data, hence a revolution at the end of an user. Furthermore, the Hypergeometric probability function is applied on the keywords and simulations are done accordingly. In this paper, we have used Hypergeometric distribution model that describes the number of successes in a sequence of n keywords to be searched from a finite population without replacement. We know that present day web searching depends on crawling and indexing techniques which relies on blindly searching for keywords without understanding their meaning. Our proposal in this paper redefines the relevance of keywords by using the syntactical characters of the keywords and applying the Hypergeometric distribution function which enables us to determine the relevance of the keywords.

## REFRENCES


[1] *A guide to Future of XML, Web Services and Knowledge Management* , by Michael C.Daconta, Leo J. Obrst, Kevin T.Smith, ISBN : 0-471-43257-1, published in 2003

[2] *A Semantic Web primer* by Grigoris Antoniou and Frank Ven Harmelen, ISBN: 978-0-262-01242-3, MIT PRESS 2008.

[3] Cosmin Striletchi, Mircea F. Vaida, *"A Web 3.0 Solution for Restraining the Web-bots Access to the Online Displayed Content"*, proc 31st International Conference on Information Technology Interfaces, IEEE 2009, ISBN: 978-953-7138-15-8, pp 633-838

[4] Dean Allemang "*Rule-based intelligence in the Semantic Web",* proc. Second International Conference on Rules, and Rule Markup Languages of the Semantic Web, IEEE Computer Society, Athens GA, Nov 2006, ISBN: 0-7695-2652-7, pp. 83-88.

[5] Li BAI, Min LIU, "*A Fuzzy –set based Semantic Similarity Matching Algorithm for Web-Services"* proc. 2008 IEEE International Conference on Services Computing. , 7-11 July 2008, pp. 529-532

[6] LI yuan, ZENG jianqiu, "*Web 3.0: A real personal web!" Beijing China , proc 3rd International Conference on Next Generation Mobile Applications, Services and Technologies.,* ISBN: 978-0-7695-3786-3, *IEEE 2009., pp. 125-128*

[7] Ora Lassila, James Hendler, *Embracing "Web 3.0" ,in IEEE Internet Computing Magazine, ISSN: 1089-7801, IEEE computer society, 2007, pp. 90-93*

[8] *Programming the Semantic Web* by Toby Segaran, Colin Evans, Jamie Taylor, ISBN: 978-0596-15381-6, O'REILLY 2009

[9] Qing Zhou, ZeQI Zheng "*An intelligent Query Expansion of Searching Related Text Information by Keywords" Proc International Conference on Web Intelligence.* 20-24 September 2004, Beijing, China. IEEE Computer Society, ISBN 0-7695-2100-2, pp 582-585.

[10] Radha Guha, "*Towards the Intelligent Web Systems"*, proc. First International Conference on Computational Intelligence, Communications Systems and Networks, , Coimbator(India), IEEE 2009, pp. 459-463

[11] Samantha K. Rajapaksha and Nuwan Kadogoda "*Internal Structure and Semantic Web Link Structure Based Ontology Ranking" , proc 4th International Conference on Information and Automation for Sustainibility,12-14 December, ISBN:* 978-1-4244-2899-1, *pp. 86-90,IEEE 2008.*

[12] Siddharth Gupta and Narina Thakur *"Semantic Query Optimisation with Ontology Simulation", International Journal of Web and Semantic Technology, Vol. 1, No.4, ISSN: 0975-9026(Online) , 2010, pp 1-10.*

[13] *XML Databases and Semantic Web* by Bhavani Thuraisingham, ISBN : 0-8493-1031-8, published 2002

[14] Yashihiro Tohma et al "*The Estimation of Parameters of the Hypergeometric Distribution and its Application to Software Reliability Growth Model "in IEEE transactions on Software Engineering Vol 17, No. 5,1991* ISSN : 0098-5589 *,pp. 483-489.*